Caractérisation de registres de langue par extraction de motifs séquentiels émergents

Jade Mekki, Nicolas Béchet, Delphine Battistelli, Gwénolé Lecorvé

Caractérisation de registres de langue par extraction de motifs séquentiels émergents

Jade Mekki[1,3], Nicolas Béchet[2], Delphine Battistelli[3], Gwénolé Lecorvé[1]

[1]Univ Rennes, CNRS, IRISA – prenom.nom@irisa.fr
[2]Univ Bretagne Sud, CNRS, IRISA – prenom.nom@irisa.fr
[3]Univ Paris Nanterre, CNRS, MODYCO – prenom.nom@parisnanterre.fr

Abstract

Language registers are the highly perceptible characteristic of written or spoken communication. In this paper we present a methodology to automatically characterize language registers using statistical tool named "emerging sequential patterns". Our approach is presented in two steps : the first one exhibits the relevance of the chosen statistical tool from artificial texts ; the second one shows that the characteristic patterns of the language registers from real data can be extracted by using this statistical tool. Experimental results show the quality of our methodology.

Keywords: Language registers, emerging sequential patterns

Résumé

Les registres de langue sont un trait saillant et très visible de la communication orale et écrite. Nous proposons dans cet article une méthodologie qui permet de caractériser automatiquement les registres de langues. Elle s'appuie sur un outil statistique particulier qui repose sur l'utilisation de motifs dits "séquentiels émergents". Les travaux que nous exposons ici présentent deux étapes : une première étape qui vérifie la pertinence de l'outil statistique choisi à partir de textes artificiels ; une seconde étape qui applique cet outil à des données textuelles réelles. Les résultats expérimentaux à partir de données réelles sont encourageants étant donnée la qualité des motifs caractéristiques des registres de langue retournés.

Mots clés : Registres de langue, motifs séquentiels émergents

1 Introduction

Le terme de registre de langue est utilisé pour rendre compte par exemple de la différence perçue entre une conversation informelle entre amis et un échange professionnel. Il peut également être utilisé pour distinguer ce qui sera perçu comme un langage soutenu en opposition notamment à un langage familier. Cette dimension langagière relève d'un ensemble de motifs linguistiques décrits comme associés de manière typique à un certain contexte de communication. Nos travaux s'intéressent à l'analyse automatique de cette dimension. Notre objectif principal est de caractériser automatiquement un registre de langue par extraction de motifs linguistiques selon une méthodologie qui consiste à ne pas poser d'*a priori* sur ces motifs qui sont envisagés à plusieurs niveaux d'abstraction de la langue (phonétique, morphosyntaxique, syntaxique et

lexical). Notre méthodologie repose plus précisément sur deux hypothèses : la première est que l'on peut décrire un registre de langue par un ensemble de motifs linguistiques ; la seconde est que les motifs séquentiels émergents constituent un outil pertinent d'extraction de ces motifs. La première hypothèse a été explorée et validée lors de travaux préliminaires que nous avons menés (Mekki, 2018) et qui nous ont permis de lister puis tester sur corpus 72 motifs linguistiques (de nature phonétique, morphosyntaxique, syntaxique et lexicale) considérés comme pertinents dans la littérature linguistique sur le sujet. L'exploration de la seconde hypothèse consiste non seulement à vérifier que l'outil des motifs séquentiels émergents permet de détecter la présence et la robustesse (quant à leur caractère discriminatoire) à une échelle importante de ces 72 motifs linguistiques mais aussi d'en détecter de nouveaux. Le présent article est plus particulièrement axé sur la question de la fiabilité de la méthode des motifs séquentiels émergents en application d'abord à des données artificielles puis à des données réelles. Plus précisément, les contributions présentées ici sont de deux ordres :

— Nous présentons ici une évaluation automatique et quantitative à partir de textes générés par des langages formels pour estimer objectivement la fiabilité des motifs extraits (lesquels sont généralement analysés manuellement quant à leur fiabilité). Ceci nous permet de démontrer la robustesse de notre outil et nous permet ensuite d'extraire des motifs linguistiques caractéristiques des registres de langue à partir de données réelles sans *a priori*.

— Les outils d'extraction classiques utilisés en linguistique de corpus dégagent généralement des motifs qui ne contiennent qu'un seul niveau d'analyse de la langue, par exemple morpho-syntaxique (ex : "*mot:il*, *mot:mange*", "*pos:pronom_personnel*, *pos:verbe*") ou syntaxique (ex : "*syntaxe:sujet*, *syntaxe:racine*"). Or, les motifs séquentiels permettent de prendre en compte plusieurs niveaux d'analyse au sein d'un seul motif, en combinant par exemple des informations morpho-syntaxiques et syntaxiques (ex : "*pos:pronom_personnel*, *syntaxe:racine*"). Ainsi, ils constituent un outil d'analyse de données textuelles puissant pour la modélisation de phénomènes linguistiques.

Après un état de l'art présenté en section 2, nous présentons notre méthodologie en Section 3. Nous exposons ensuite plusieurs expériences en Section 4 qui permettent de valider l'hypothèse et donc la pertinence de notre méthodologie.

2 État de l'Art & Positionnement

Toute production langagière est évaluée par l'interlocuteur. Il la situe dans un "registre", qui peut être vu intuitivement comme un certain usage de la langue à un moment donné et/ou dans un contexte donné. Cette notion se trouve abordée dans des travaux divers en linguistique comme en sociolinguistique. (Ferguson, 1982) définit les registres comme une variation *"dans laquelle la structure linguistique varie en fonction des occasions d'utilisation"*. (Ure, 1982) associe cette variation aux activités humaines : *"chaque communauté linguistique a son propre système de registres... correspondant à l'éventail des activités que ses membres exercent normalement"*. Selon l'angle d'étude privilégié, on observe dans la littérature linguistique diverses manières de partitionner l'espace linguistique en différents registres. Par exemple, (Ilmola, 2012) propose de distinguer les registres familier, populaire et vulgaire dans des journaux satiriques, là où (Borzeix et Fraenkel, 2005) catégorisent différentes situations de communication au travail en

opposant, par exemple, *"la communication fonctionnelle"* à *"la communication relationnelle"*. Il apparaît rapidement une difficulté définitoire et terminologique dans les travaux abordant cette notion. Ainsi, les dénominations "niveau de langue" (Gadet, 1996) ou encore "genre" co-existent avec celle de "registre" (Biber, 2019) . L'état de l'art fait par (Argamon, 2019) montre que les travaux consacrés à l'analyse automatique de cette dimension ne recourent que de manière très marginale au terme "registre" et utilisent préférentiellement celui de "style", de "genre" ou encore de "(degré de) formalité". Dans le contexte du TAL (au sens strict du terme) on ne trouve de fait, à notre connaissance, aucune étude qui utilise le terme de "registre". On relève pourtant des approches qui s'y intéressent puisqu'elles traitent de la question du degré de formalité d'une phrase (Sheikha et Inkpen, 2010) ou d'un document (Pavlick et Tetreault, 2016). D'autres approches s'intéressent au style d'un texte, au travers de la problématique de l'attribution automatique d'auteur. (Stamatatos, 2009) propose un état des lieux de cette problématique qui peut être explorée dans des contextes très différents (billets de blogs (Schler et al., 2006), messages textuels (sms) (Cougnon et Fairon, 2014), ou bien textes anonymes (Eisenstein, 2013)). Comme le rappelle (Stamatatos, 2009), le style d'un auteur est le résultat de différents choix à plusieurs niveaux d'analyse de la langue. Le plus évident et le plus étudié est le niveau lexical (analyse de la longueur des mots, de la longueur des phrases dans un texte, de la richesse lexicale ou bien de la fréquence de n-grammes de mots par exemple sont classiques dans ce domaine). Pour (Argamon et al., 2007), il est communément accepté par ailleurs que les mots grammaticaux (tels que les prépositions, les déterminants, les auxiliaires, les temps verbaux modaux, *etc*.) sont intéressants à prendre en compte pour l'étude de la dimension stylistique tandis que d'autres (tels que les noms ou adjectifs) ne le sont pas. Les caractéristiques morphosyntaxiques et syntaxiques sont également largement utilisées pour caractériser le style (Sidorov et al., 2014). Enfin, d'autres études se sont concentrées sur les informations graphiques en se basant sur des n-grammes de caractères, des types des graphèmes (lettres, nombres, ponctuation, majuscules, *etc*.). Tous ces travaux de TAL mettent en exergue l'importance de la diversité des niveaux d'abstraction de la langue à prendre en compte pour travailler sur l'identification des styles d'auteurs.

Si nous avons trouvé peu de travaux en TAL sur les registres de langue en tant que tels, nous avons relevé l'existence de nombreux travaux du côté de la linguistique de corpus qui utilisent cette fois explicitement le terme de "registre". Ce terme est par exemple utilisé par Biber depuis ses premiers travaux (Biber, 1991) jusqu'à aujourd'hui (Biber et Conrad 2019). Dans ses travaux récents, Biber définit un registre comme *"une variété linguistique associée à une situation particulière d'utilisation (en comprenant des buts particuliers de communication)"* (Biber et Conrad, 2019). L'identification d'un registre repose sur des *"descripteurs linguistiques qui ont toujours des rôles fonctionnels"* (Biber et Conrad, 2019), c'est à dire qu'ils sont choisis selon le contexte et l'objectif de la communication. Le style se différencie, selon Biber, du registre dans la mesure où les descripteurs linguistiques ne sont dans ce cas pas fonctionnels car ils reflètent *"plutôt des préférences esthétiques, associées à des auteurs particuliers ou des périodes historiques"* (ibid.). D'un point de vue méthodologique, (Poudat et Landragin, 2017) pointe certaines limites à l'approche de Biber. La première est que *"le corpus doit d'abord faire système pour le chercheur, qu'il soit supposément homogène ou au contraire structuré suivant une hypothèse de variété"*. De manière similaire les descripteurs relevés *"doivent être sinon réfléchis, du moins sélectionnés dans le cadre d'hypothèses linguistiques ou interprétatives spécifiques et*

explicites". Or, certains auteurs (Branca-Rosoff, 1999 ; Poudat et Landragin, 2017) mettent en exergue l'absence de justification quant à la sélection de tel ou tel descripteur par Biber. Une manière pour nous de répondre à ces limites consiste à proposer une méthodologie fondée sur l'extraction de motifs séquentiels sans *a priori*. Nous appelons *registre de langue l'utilisation d'un ensemble de motifs linguistiques spécifiques à un contexte de communication en ne prenant pas en considération le principe de fonctionnalité associé aux descripteurs linguistiques*. En cela, nous nous éloignons de la notion de *"registre"* comme définie dans (Biber et Conrad, 2019). Nous préférons utiliser le terme *"registre"* afin d'éviter toute notion hiérarchique ou bien normative qui pourrait se refléter à travers l'expression *"niveau de langue"* par exemple. Notre étude partitionne l'espace linguistique en trois registres principaux : familier, courant, soutenu. Bien que nous admettions sans difficulté qu'il existe un continuum entre ces trois registres, cette partition découle du besoin d'un découpage en valeurs discrètes pour un traitement automatique. Nous utilisons les motifs séquentiels émergents comme outil automatique puisque ces derniers nous permettent de garder une notion d'ordre entre les objets linguistiques grâce aux motifs séquentiels et de traiter plusieurs niveaux d'analyse de la langue grâce aux itemsets. La difficulté liée à cet outil réside dans l'évaluation des motifs retournés : comment savoir si ces derniers sont pertinents ? Notre contribution se trouve dans la proposition d'une méthodologie robuste qui extrait sans *a priori* des motifs caractéristiques des registres de langue. Cette méthodologie est validée par deux expérimentations différentes : la première à partir de textes artificiels afin d'évaluer la solidité de l'outil d'extraction, la seconde à partir de données réelles afin de confirmer les motifs listés dans la littérature scientifique sur le sujet et mettre à jour de nouveaux descripteurs.

3 Méthodologie

La difficulté majeure des outils d'extraction de motifs séquentiels émergents réside dans le fait que les motifs extraits doivent être évalués et analysés manuellement en vue de vérifier leur fiabilité et pertinence, comme dans (Legallois et al., 2016) par exemple. Pour remédier à cela nous avons décidé de mettre en place une méthodologie qui évalue automatiquement et quantitativement cet outil statistique en partant de données artificielles. Dans cette section, nous exposons la génération de textes artificiels par l'utilisation de langages formels (Section 3.1) avant d'introduire les techniques de fouille de données sur lesquelles notre approche s'appuie : les motifs séquentiels fréquents, clos et émergents (Section 3.2). Enfin nous présentons la notion de motifs séquentiels émergents (Section 3.3).

3.1 Langages formels

Les grammaires utilisées pour générer les textes artificiels sont des grammaires hors-contexte probabilistes. Elles peuvent être définies par un cinq-uplet $< N, T, R, S, P >$ où N est l'ensemble des symboles non-terminaux, T est l'ensemble des symboles terminaux, R est l'ensemble des règles r_i de la forme $A \rightarrow \beta$, S est l'axiome de départ, P est l'ensemble des probabilités p_i associées aux règles r_i telles que $\sum_\beta \Pr(A \rightarrow \beta) = 1, \forall A \in N$. Voici un exemple pour générer les énoncés suivants : *"le chat dort", "un chat dort", "le chat joue", "un chat joue"*.

$$
\begin{array}{rcl@{\qquad}rcl}
S & \longrightarrow & (SN + SV)_1 & VB & \longrightarrow & "dort"_{0,5} \mid "joue"_{0,5} \\
SN & \longrightarrow & (DET + NC)_1 & DET & \longrightarrow & "le"_{0,5} \mid "un"_{0,5} \\
SV & \longrightarrow & (VB)_1 & NC & \longrightarrow & "chat"_1
\end{array}
$$

3.2 Fouille de motifs séquentiels

La fouille de motifs séquentiels introduite par (Agrawal et al., 1995) permet d'identifier des régularités qui considèrent la temporalité dans des bases de données. Ce que nous appelons motifs séquentiels est un sous-ensemble d'une séquence. Un $itemset$ noté I, est composé d'un ensemble de littéraux appelés $item$ noté i. Un itemset est donc représenté par $I = (i_1, i_2, ...i_n)$. Une *séquence* S est une liste ordonnée d'itemsets et est notée $S = \langle I_1 \dots I_m \rangle$. Par exemple, la séquence $\langle (a, b, c)(a, d)(a, b) \rangle$ est une séquence de trois itemsets chacun composé respectivement de trois, deux et deux items. Une séquence $S_1 = \langle I_1, I_2, ...I_n \rangle$ est une *sous séquence* de $S_2 = \langle I'_1, I'_2, ...I'_m \rangle$ s'il existe des entiers $1 \leq j_1 < \dots < j_n \leq m$ tels que $I_1 \subseteq I'_{j1}, \dots, I_n \subseteq I'_{jn}$. Ainsi, S_1 est une sous séquence de S_2 notée $S_1 \leqslant S_2$, par exemple : $\langle (a)(d) \rangle \leqslant \langle (a, b, c)(a, d)(a, b) \rangle$. Une base de données séquentielles (ou *Sequential DataBase*, noté SDB) est un ensemble de tuples noté (sid, S), où sid est un identifiant de séquence et S une séquence. La table 1 ci-dessous représente une base de données séquentielles de quatre séquences.

Identifiant de séquence	Séquence
1	$\langle (a, b, c)(a, d)(a, b) \rangle$
2	$\langle (d)(a, d)(e) \rangle$
3	$\langle (a)(a, d)(b, c) \rangle$
4	$\langle (b, c)(a, d)(c) \rangle$

TABLE *1 – Exemple de base de données séquentielles notée* SDB_{ex}

Motifs séquentiels fréquents Le *support absolu* d'une séquence S_1 dans une base de données SDB, noté $sup_{SDB}(S_1)$, est le nombre de tuples contenant S_1 dans la base SDB. Par exemple, le motif $S_1 = \langle (a)(a) \rangle$ dans la base SDB a pour support absolu $sup_{SDB}(S_1) = 2$: les séquences 1 et 2 contiennent un itemset avec a suivi d'un itemset avec a. Le *support relatif* d'une séquence dans SDB est le *support absolu* divisé par le nombre total de séquences présentes dans SDB : $sup_{SDB}(S_1) = \frac{|\{(sid,S)|(sid,S) \in SDB \wedge (S_1 \preceq S)\}|}{|SDB|}$. Un motif est dit *fréquent* lorsque son support est supérieur ou égal à un seuil fixé par l'utilisateur appelé *support minimum* ou *minsup*. Un algorithme de fouille de motifs séquentiels a pour but d'extraire tous les motifs fréquents dans une base de données : tous les motifs dont le support est supérieur ou égal au seuil *minsup*. Toutefois, les motifs extraits peuvent être très (trop) nombreux et redondants. Afin d'éviter cela, il existe une représentation condensée sans perte d'information : *les motifs séquentiels clos*.

Motifs séquentiels clos Les *motifs séquentiels clos* sont introduits par (Yan et al., 2003). Un motif fréquent S est dit clos s'il n'existe aucun motif fréquent S' tel que $S \leqslant S'$ et $sup(S) = sup(S')$. Par exemple, le motif $S_1 = \langle (a) \rangle$ n'est pas clos puisqu'il existe un motif $S'_1 = \langle (a, d) \rangle$

tel que $S_1 \leqslant S_1^{'}$ et $sup_{SDB}(S_1) = sup_{SDB}(S_1^{'})$. En revanche, le motif $S_2 = \langle (a,d)(e) \rangle$ est clos car il n'est inclus dans aucun motif fréquent S' ayant le même support.

Contraintes d'extraction de motifs séquentiels Afin de limiter le nombre de motifs extraits, il peut-être pertinent de mettre en oeuvre des contraintes (Dong et Pei, 2007). Deux contraintes sont généralement employées : la contrainte de fréquence minimum avec le seuil *minsup* tel que présenté au paragraphe *Motifs sequentiels fréquents*, et la contrainte de *gap*. Un motif avec un *gap* $[M, N]$, noté $S_{[M,N]}$, est un motif dont chaque couple d'itemsets est séparé par au moins $M-1$ itemsets et au plus $N-1$ itemsets. Par exemple, $S_{[1,3]} = \langle (d)(a) \rangle$ est un motif qui apparaît dans les séquences 1 et 2.

3.3 Motifs séquentiels émergents

Les motifs séquentiels émergents sont des séquences dont le support augmente de manière significative d'un ensemble de données à un autre. Le taux de croissance d'un motif S, noté $GrowthRate(S)$, est le rapport des supports d'un même motif dans deux ensembles de données différents (R_1, R_2) (Equation 1). Un motif est dit $emergent$ si son $GrowthRate$ est supérieur à un seuil fixé par l'utilisateur : $threshold$.

$$GrowthRate(S_{R_1|R_2}) = \begin{cases} \infty, & si \sup_{R_2}(S) = 0 \\ \frac{\sup_{R_1}(S)}{\sup_{R_2}(S)}, & sinon \end{cases} \quad (1)$$

4 Expérimentations

La difficulté des motifs séquentiels émergents réside dans l'interprétation des motifs retournés et de leur fiabilité. Nous proposons un protocole expérimental qui permet de répondre à ces difficultés en testant la robustesse des motifs séquentiels émergents pour la caractérisation des registres de langue à partir de textes artificiels. Ainsi, nous procédons à deux expériences : la première a pour but de valider les motifs séquentiels émergents comme pertinents pour caractériser un registre de langue à partir de textes artificiels (Sous-section 4.1), la seconde extrait les motifs séquentiels à partir de données réelles en considérant les motifs extraits comme fiables (Sous-section 4.2).

4.1 Expériences à partir de données artificielles

Les textes artificiels sont utilisés afin de constituer un corpus dans lequel nous connaissons les descripteurs présents ainsi que leurs proportions puisque les grammaires hors-contexte probabilistes nous permettent d'insérer des traits linguistiques plus ou moins fréquemment. Savoir *a priori* quels sont les motifs caractéristiques d'un registre et à quelles fréquences nous permet d'évaluer la fiabilité de l'extraction automatique des ces derniers. Ainsi, nous cherchons à savoir si :

— Les motifs que nous savons caractéristiques d'un registre sont effectivement extraits par l'algorithme d'extraction de motifs séquentiels émergents ;

— Réciproquement, les motifs extraits dont le taux de croissance est supérieur à 1 contribuent bien au registre que nous voulons caractériser et qu'inversement les motifs dont le taux de croissance est inférieur ou égal à 1 ne contribuent pas au registre.

Corpus Le corpus a été généré à partir de grammaires hors-contexte probabilistes qui nous ont permis d'implémenter des traits linguistiques avec des pondérations variables dans les registres différents. Au total, elles sont composées d'un ensemble N qui comprend 22 symboles non terminaux (ex : "DET","NC"), d'un ensemble T de 36 symboles terminaux (ex : "le","$chat$") et enfin d'un ensemble R de 51 règles r_i (ex : $SN \longrightarrow DET + NC$) associées à 51 probabilités p_i (ex : $DET \longrightarrow "le"_{0.50}|"un"_{0.50}$). Deux grammaires génératives hors-contexte sont écrites pour deux registres : familier et soutenu. Elles nous permettent de pondérer des règles contextuelles grâce aux probabilités p_i qui changent de valeurs selon le registre. Nous avons donc pu introduire des motifs linguistiques caractéristiques de ces deux registres dans chaque grammaire. Ces motifs linguistiques sont issus d'une étude préliminaire (Mekki et al., 2018) et se situent à plusieurs niveaux d'abstraction de la langue. Nous donnons quelques exemples de ces derniers pour le registre familier : det+nc, (on | ça)+vb, ø...pas, vb+sj+?, (radoter | chanter), (chanson | chansonnette | musique). Tandis que pour le soutenu, ils sont : det+adj+nc, (il | elle)+vb, ne...pas, sj+vb+?, (chanter | répéter), (romance | ballade). Grâce à ces grammaires génératives, des textes de taille variable ont été créés pour les deux registres (100, 1000 et 10000 séquences). Nous avons choisi de travailler sur un corpus écrit de 1000 phrases. Chaque mot de ce corpus est étiqueté avec son lemme, sa catégorie morphosyntaxique et sa fonction syntaxique. Nous segmentons le corpus au niveau de la phrase malgré les limites attachées à ce type de décision qui repose sur une ponctuation considérée comme déterminante (des cas des subordonnées séparées de la principale par un point, ou encore des textes non ponctués posent évidemment question avec ce type de décision). En cela, nous nous rallions à l'idée de (Gautier, 2014) pour qui une phrase dans un corpus composé de textes écrits est un segment graphique qui impact l'interprétation du lecteur. En effet, la ponctuation forte a "*un rôle opérateur*" (Gautier, 2014) qui déclenche une opération cognitive appelée "*wrap-up effect*" (Charolles et Lamiroy, 2001) : le lecteur fait une mise à jour du modèle discursif et se représente les données verbales d'une manière plus condensée. Ainsi chaque corpus est segmenté à l'échelle de la phrase afin de composer une base de donnée séquentielle où chaque séquence représente une phrase.

Extraction des motifs séquentiels émergents Dans ce paragraphe nous présentons les éléments et paramètres choisis pour l'extraction des motifs séquentiels émergents. Nous avons réalisé deux extractions : motifs fréquents du premier registre par rapport aux motifs fréquents du second registre puis motifs clos du premier registre par rapport aux motifs fréquents du second registre. Les paramètres fixés pour ces deux extractions sont les suivants : les registres caractérisés sont le familier noté R_1 et le soutenu noté R_2 ; le nombre de séquences est de 1000 ; l'algorithme utilisé pour l'extraction des motifs fréquents et clos est CloSpec (Béchet et al., 2015) ; le $Minsup_1$ pour l'extraction des motifs fréquents et clos de R_1 est de 5% ; et le $Minsup_2$ pour l'extraction des motifs fréquents et clos de R_2 est de 2,5% ; le seuil n'est pas fixé afin d'obtenir l'ensemble des motifs ; enfin la contrainte de *gap* est de $P[1, 1]$ (les motifs sont donc contigus). Nous précisons que pour des raisons de complexité algorithmique nous

devons fixer deux $minsup$ différents : le premier est le $minsup$ qui filtre les motifs de R_1 que nous souhaitons caractériser, le second $minsup$ filtre les motifs de R_2 par rapport auxquels nous caractérisons R_1. Ici, $Minsup_2$ représente la moitié de $Minsup_1$ afin d'assurer que les motifs comparés soient au minimum deux fois moins présents dans le registre R_2. Idéalement ce second $minsup$ devrait être égal à 1 en valeur absolue afin de comparer les motifs de R_1 à tous les motifs de R_2 quelque soit leur fréquence mais ceci n'a pas pu être appliqué pour cause de complexité algorithmique trop élevée. Tous les motifs sont extraits et nous calculons un taux de croissance pour chacun d'entre eux car notre hypothèse de départ est que les motifs séquentiels émergents sont pertinents pour caractériser un registre de langue. Aussi, si nous trions les motifs de manière décroissante par rapport à leur $GrowthRate$, alors tous les motifs qui contribuent au registre devraient être en tête et ceux qui n'y contribuent pas en queue. Pour vérifier cela, nous ne fixons pas de $threshold$ afin d'obtenir tous les motifs quelque soit leur $GrowthRate$: nous vérifions alors que les motifs dont le $GrowthRate$ est inférieur ou égal à 1 ne sont pas des motifs qui contribuent au registre que nous souhaitons caractériser. À terme, un $threshold$ sera proposé afin d'obtenir uniquement les motifs séquentiels émergents pertinents.

Protocole d'évaluation Afin d'évaluer les motifs retournés nous devons définir deux éléments : comment labelliser les motifs et quelles métriques utiliser pour les évaluer. L'utilisation de textes artificiels nous permet de connaître *a priori* les motifs linguistiques caractéristiques d'un registre dans le but de labelliser les motifs comme "bon" ou "mauvais". Nous cherchons simplement si les motifs introduits dans les grammaires sont bien présents dans les motifs retournés comme émergents, c'est à dire dont le taux de croissance est supérieur à 1. Ainsi pour labelliser un motif : comme vrai nous vérifions la présence d'un motif linguistique attendu pour le registre familier, comme faux nous vérifions l'absence d'un motif linguistique pour le registre familier. L'évaluation doit prendre en compte deux points : le motif doit caractériser un registre et le classement des motifs obtenu en les triant par taux de croissance décroissant doit mettre en tête tous les motifs évalués comme "bon". Ainsi, nous avons utilisé des métriques issues du domaine de la recherche d'information afin de mesurer la qualité du classement des motifs extraits et leurs pertinences : Area Under Receiver Operating Characteristic (AUROC) (Narkhede, 2018), Average precision (AP) (Kishida, 2005) et Normalized Discounted Cumulative Gain (NDCG) (McSherry et Najork, 2008).

Résultats expérimentaux Les résultats sont présentés table 2 Nous pouvons expliquer les très hauts scores de l'AP et de l'NDCG par le fait qu'ils soient lissés par la moyenne des résultats.

Couple de registres	Extraction	AP	AUROC	NDCG
familer X soutenu	Freq X freq	0.995	0.865	0.999
	Clos X freq	0.953	0.908	0.993
soutenu X familier	Freq X freq	0.999	0.947	0.999
	Clos X freq	0.995	0.960	0.999

TABLE 2 – *Résultats des extractions de motifs séquentiels émergents : familier par rapport soutenu*

Ces valeurs nous permettent de valider notre hypothèse selon laquelle les motifs séquentiels

émergents sont pertinents pour caractériser un registre de langue. L'indicateur $GrowthRate$ est donc un indicateur robuste auquel nous pouvons nous fier.

4.2 Expériences à partir de données réelles

Corpus Pour le corpus nous avons utilisé le modèle proposé par (Lecorvé et al., 2018), c'est à dire un classifieur semi-supervisé qui prédit le registre d'un texte donné : il apprend à partir d'une graine annotée manuellement puis itérativement rajoute des textes tirés d'un ensemble de pages récoltées du web à partir de requêtes composées de lexiques familiers et soutenus. Le corpus annoté se compose d'un ensemble de 113 027 séquences pour le familier, 331 740 pour le courant et 128 866 pour le soutenu.

Extraction des motifs séquentiels émergents Les motifs clos ont l'avantage de retourner des motifs qui contiennent plus d'itemsets que les motifs fréquents grâce à la notion de clôture. Cela nous permet de réduire le nombre de motifs sans perte d'information, mais également de retourner des motifs plus facilement interprétables. Les résultats du paragraphe *Résultats expérimentaux* montrent que les motifs fréquents et clos ont tous les deux de bons scores sans différence notable entre eux. Les motifs clos sont généralement plus longs et sont donc plus intelligibles : c'est pourquoi nous privilégions les motifs clos pour l'extraction de motifs séquentiels émergents à partir de données réelles. Les deux registres considérés sont également le familier et le soutenu. Les différents paramètres fixés pour ces différentes extractions sont les mêmes que ceux utilisés pour l'extraction à partir de textes artificiels.

Résultats du familier par rapport au soutenu La table 3 présente différents motifs séquentiels émergents (tous les exemples viennent de notre corpus). Les motifs de 1 à 4 sont intéressants puisqu'ils confirment les motifs identifiés dans la littérature scientifique comme spécifiques du registre familier. Le motif 1 renvoie à l'absence de la double négation (Bilger et Cappeau, 2004), le motif 2 illustre la contraction du syntagme "cela est" (Golubéva-Monatkina, 1991), le motif 3 donne un exemple de la répétition des signes de ponctuation (Branca-Rosoff, 1999) et le motif 4 avec la contraction du "nous" en "on" (Bilger et Cappeau, 2004). Ces résultats permettent de confirmer des descripteurs listés dans la littérature scientifique empiriquement admis comme caractéristiques de tel ou tel registre. L'extraction de ces descripteurs connus sans *a priori* à partir d'un large corpus issu du web permet de confirmer et de justifier leur caractère discriminant de manière automatique et déductive. En outre, de voir émerger des motifs séquentiels connus de la littérature scientifique linguistique renforce notre confiance en la fiabilité des nouveaux motifs séquentiels émergents qui ne sont pas encore identifiés comme caractéristiques d'un registre tels que les motifs 5 à 9 présentés table 3. Le motif 5 pourrait être lié aux usages d'écriture numérique comme lorsque nous relançons l'interlocuteur ou bien avec un terme ponctuant, par exemple : *"Tu l'as bien là, non ?", "Et les clés de la tire, dis ?", "alors, yes or no ?"*. Le motif 6 indique que les constructions verbales pronominales seraient caractéristiques du familier, par exemple : *"Elle se coltine une bouille d'épagneul harassé", "Une jeune femme se pointe bientôt, avec des bières." "et même demander aux chinois de se magner à fabriquer des nounours"*. Le motif 7 est une sur représentation des expressions multimots utilisées en tant que nom propre telles que *"mézigue Bibi"*. Le motif 8 montre l'utilisation

	Motif	**Exemples**
	Familier vs. soutenu	
1	⟨ (pos:*auxiliaire*) , (syntax:*advmod*, pos:*adverbe*, lemme:*pas*) ⟩	• *"Hé ! dis, vieux, je l'ai pas refroidie, au moins ?"* • *"c'est pas non plus ton frometon à toi, béby !"*
2	⟨ (lemme:*c*) , (pos:*ponctuation*, *'*, lemme:*"*, syntax:*ponctuation*) , (lemme:*etre*, syntax:*cop*) ⟩	• *"c'est pas reluisant"* • *"c'est chié la vie avec toi !"* • *"Pffff. C'était même pas vraix."*
3	⟨ (pos:*ponctuation*, syntax:*ponctuation*) , (pos:*ponctuation*) , (pos:*ponctuation*) ⟩	• *"Et c'est 80 euros d'ailleurs (... ahahahaha)"* • *"ne le laissont pas filer ! ! !"*
4	⟨ (syntax:*nsubj*, lemme:*on*) ⟩	• *"on l'a jamais vu s'afficher avec des meufs"*
5	⟨ (pos:*poncutation*, mot:*?*, lemme:*?*) ⟩	• *"ça compense un manque ou quoi ?"*
6	⟨ (pos:*pronom*, mot:*se*) , (pos:*verbe*) ⟩	• *"pour pas se faire chopper"*
7	⟨ (pos:*pronom_personnel*, syntaxe:*expression_multimots*) ⟩	• *"le Tombeur de Saint-Cloud"* • *"miss Zouzou"*
8	⟨ (syntax:*auxiliaire*) , (pos:*adverbe*) ⟩	• *"C'est bien. Ouais."*
9	⟨ (pos:*verbe*) , (pos:*adverbe*, syntaxe:*modifieur*) , (pos:*adverbe*) ⟩	• *"ça se passera très bien"* • *"où ça se finit pas hyper bien"*
	Soutenu vs. familier	
10	⟨ (lemme:*ne*, pos:*adverbe*) , (pos:*verbe*) ⟩	• *"ne valait-il pas mieux"*
11	⟨ (pos:*pronom*, mot:*me*, lemme:*me*) ⟩	• *"il me semblait"*
12	⟨ (pos:*adverbe*, mot:*vous*, lemme:*vous*) ⟩	• *"vous qui l'aimiez tant"*
13	⟨ (pos:*ponctuation*, mot:*;* , lemme:*;*) ⟩	• *"du Venezuela et du Panamá ; enfin, le Brésil"*
14	⟨ (mot:*comme*, lemme:*comme*) ⟩	• *"comme elle n'avait guère"*

TABLE *3 – Résultats des extractions de motifs séquentiels émergents : Familier vs. soutenu et Soutenu vs. familier*

plus fréquente pour le familier d'un verbe auxiliaire suivi d'un adverbe, par exemple : *"Il est vachement crayeux de teint, le défunt."*, *"C'est mal foutu cette affaire..."*, *"elle a pleuré super fort"*. Enfin, le motif 9 marque l'enchaînement d'un verbe et de deux adverbes, par exemple : *"Il pige très bien"*, *"Ça me fait hyper mal."*, *"c'est vachement bien"*.

Résultats de la caractérisation du soutenu par rapport au familier Le motif 10 de la table 3 confirme la pertinence de la négation pour caractériser un registre de langue puisqu'il présente sa forme non contractée tandis que sa forme contractée est caractéristique du familier, par exemple : *"je savais que je ne la quitterais plus, tout aussi bien que je savais que je ne me mettrais plus à travailler"*. Les motifs 11 et 12 montrent l'importance des pronoms personnels avec notamment l'utilisation du pronom "vous" au détriment du pronom "tu" (Bilger et Cappeau, 2004), par exemple : *"Voulez-vous auparavant voir votre mère une dernière fois ?","Je ne vous savais pas ce don de sarcasme aiguisé."*. Les motifs 13 et 14 semblent indiquer des constructions de phrases complexes avec le signe de ponctuation ";" et le comparateur "comme" qui introduit des comparaisons voire des métaphores, par exemple : *"J'ai été signalé comme saint-simonien et j'ai failli être tué"*, *"son arrivée fit éclater mes sanglots, comme à un enterrement"*. En outre, beaucoup de motifs issus de la littérature linguistique sur le registre soutenu se fondent

sur les temps verbaux. Or nous n'avons pas annoté assez finement notre corpus pour cela, nous le ferons dans de futurs travaux.

5 Conclusion

Dans cet article nous avons proposé une méthodologie qui permet de valider la pertinence de l'utilisation de motifs séquentiels émergents afin de caractériser des registres de langue en français. L'intuition derrière notre proposition est que les motifs émergents d'un registre par rapport à un autre permettront de mettre en avant ses caractéristiques. Une première expérimentation utilisant un corpus à base de données artificielles a permis de montrer la fiabilité de l'outil pour cette tâche. Les résultats de la seconde expérimentation à base de données réelles ont confirmé et justifié certaines hypothèses de la littérature concernant les registres de langues en français, nous encourageant à poursuivre l'exploration des registres avec cette approche. Nous aimerions désormais varier les expériences à partir de données réelles en testant des valeurs de *gap* différentes et ainsi trouver des motifs non contigus. Nous voudrions également extraire des motifs à partir d'un corpus plus volumineux et introduire d'autres traits tels que les temps verbaux, la morphologie d'un mot, etc... On trouve ici tout l'intérêt d'une approche qui exploite à l'échelle d'un seul motif des éléments de tous les niveaux d'analyse de la langue.

Remerciements

Ce travail a bénéficié du soutien du projet TREMoLo [1] (ANR-16-CE23-0019) de l'Agence Nationale de la Recherche (ANR).

References

Agrawal R., Srikant R. (1995). "Mining sequential pattern". *icde*. Vol. 95, pp, 3–14.

Argamon S. (2019). "Register in computational language research". *Regiter Studies* 1.1, pp. 100–135.

Argamon S., Whitelaw C., Chase P., Hota S. R., Garg N., Levitan S. (2007). "Stylistic text classification using functional lexical features". *Journal of the Association for Information Science and Technology* 58.6.

Béchet N., Cellier P., Charnois T., Crémilleux B. (2015). "Sequence mining under multiple constraints", *Proceedings of the 30th Annual ACM Symposium on Applied Computing.* ACM, pp. 908–914.

Biber D. (1991). *Variation across speech and writing*, Cambridge University Press.

Biber D., Conrad S. (2019). *Register, genre, and style*. Cambridge University Press.

Bilger M., Cappeau P. (2004). "L'oral ou la multiplication des styles". *Langage et société* 3, pp. 13–30.

Borzeix A., Fraenkel B. (2005). "Langage et travail(communication, cognition, action)." *CNRS communication*.

Branca-Rosoff S. (1999). "Des innovations et des fonctionnements de langue rapportés à des genres." *Langage & société* 87.1, pp. 115–129.

Legallois D., Charnois T., Poibeau T. (2016). "Repérer les clichés dans les romans sentimentaux grâce à la méthode des «motifs»." *Lidil. Revue de linguistique et de didactique des langues*, (53), 95-117.

1. https://tremolo.irisa.fr/

Charolles M., Lamiroy B. (2001). “Syntaxe phrastique et transphrastique du but au résultat.” *Macro-syntaxe et macrosémantique, Actes du colloque international dAarhus, 17-19 mai 2001*, Peter Lang, pp. 383–419.

Cougnon L. A., Fairon C. (2014). *SMS Communication : A linguistic approach.* Vol. 61. John Benjamins Publishing Company.

Dong G., Jian P. (2007). *Sequence data mining.* Vol. 33. Springer Science Business Media.

Eisenstein J. (2013). “What to do about bad language on the internet.” *Proceedings of HLT-NAACL.*

Ferguson C. (1982). “Simplified registers and linguistic theory”. *Exceptional language and linguistics*, pp. 49–66.

Gadet F. (1996). “Niveaux de langue et variation intrinsèque”. *Palimpsestes. Revue de traduction* 10, pp. 17–40.

Gautier A. (2014). “Phrase et syntaxe : sur quelques aspects de l’intégration”. *Langue française* 2, pp. 27–41.

Golubéva-Monatkina N. (1991). “Gadet, Françoise. Le français ordinaire. Paris : Armand Colin, 1989”. *Canadian Modern Language Review* 47.4, pp. 800–802.

Ilmola M. (2012). *Les registres familier, populaire et vulgaire dans le canard enchaıné et charlie hebdo : étude comparative*.

Kishida K. (2005). *Property of average precision and its generalization : An examination of evaluation indicator for information retrieval experiments.* National Institute of Informatics Tokyo, Japan.

Lecorvé G., Ayats H., Fournier B., Mekki J., Chevelu J., Battistelli D., Béchet N. (2018). “Construction conjointe d’un corpus et d’un classifieur pour les registres de langue en français”.

McSherry F., Najork M. (2008). “Computing information retrieval performance measures efficiently in the presence of tied scores”. *European conference on information retrieval*. Springer, pp. 414–421.

Mekki, J., Battistelli, D., Lecorvé, G., Béchet, N. (2018). Identification de descripteurs pour la caractérisation de registres.

Narkhede S. (2018). “Understanding AUC-ROC Curve”. *Towards Data Science* 26.

Pavlick E., Tetreault J. (2016). “An empirical analysis of formality in online communication”. *Transactions of the Association of Computational Linguistics* 4.1.

Poudat C., Landragin F. (2017). *Explorer un corpus textuel : Méthodes-pratiques-outils.* De Boeck Superieur.

Schler J., Koppel M., Argamon S., & Pennebaker J. W. (2006). “Effects of Age and Gender on Blogging.” *Proceedings of the AAAI spring symposium : Computational approaches to analyzing weblogs*. Vol. 6.

Sheikha F. A., Inkpen D. (2010). “Automatic classification of documents by formality”. *IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*.

Sidorov G., Velasquez F., Stamatatos E., Gelbukh A., & Chanona-Hernández L.l. (2014). “Syntactic n-grams as machine learning features for natural language processing”. *Expert Systems with Applications* 41.3.

Stamatatos, E. (2009). “A survey of modern authorship attribution methods”. *Journal of the American Society for information Science and Technology* 60.3, pp. 538–556.

Ure, J. (1982). "Introduction : approaches to the study of register range". *International Journal of the Sociology of Language*, 1982(35), 5-24.

Yan X., Han J., & Afshar R. (2003). “CloSpan : Mining : Closed sequential patterns in large datasets”. *Proceedings of the 2003 SIAM*. SIAM, pp. 166–177.